PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2000-035965
(43)Date of publication of application : 02.02.2000

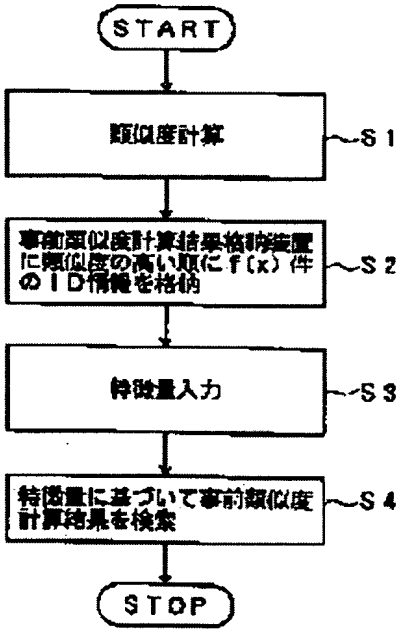
(51)Int.Cl. G06F 17/30
G06T 7/00
// G06T 1/00

(21)Application number : 10-203583
(22)Date of filing : 17.07.1998

(71)Applicant : NIPPON TELEGR & TELEPH CORP <NTT>
(72)Inventor : AKAMA HIROKI
SATO MICHIYOSHI
MITSUI KAZUYOSHI
KUSHIMA KAZUHIKO

(54) METHOD AND DEVICE FOR RETRIEVING SIMILAR FEATURE QUANTITY AND STORAGE MEDIUM STORING RETRIEVAL PROGRAM OF SIMILAR FEATURE QUANTITY

(57)Abstract:
PROBLEM TO BE SOLVED: To perform fast retrieval by retrieving a pre-similarity calculation result storage device based on retrieval key feature quantity when feature quantity in a database is given as the retrieval key feature quantity and returning a pre-similarity calculation result as a retrieval result.
SOLUTION: The whole feature quantities are preliminarily made keys, similar calculation in a database is performed and other feature quantities and similarity are calculated (S1). ID information for upper rank f(x) matters is stored in a pre-similarity calculation result storage device in order of similarity with a similarity sequence or together with similarity value as occasion demands (S2). And, when feature quantity in the database is given as retrieval key feature quantity (S3), the pre-similarity calculation result storage device is retrieved based on the retrieval key feature quantity and a pre-similarity calculation result is returned as a retrieval result (S4). According to this method, it is possible to very fast perform retrieval even if the number of the entire database feature quantity data pieces is large.



LEGAL STATUS

[Date of request for examination] 21.11.2000
[Date of sending the examiner's decision of rejection] 25.11.2003
[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]
[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

Citation 1

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（A）　(11)特許出願公開番号
特開2000－35965
（Ｐ２０００－３５９６５Ａ）
(43)公開日　平成12年２月２日(2000.2.2)

(51)Int.Cl.[7]	識別記号	FI			テーマコード（参考）
G06F 17/30		G06F 15/40	370	G	5B050
G06T 7/00		15/401	310	D	5B075
// G06T 1/00		15/403	350	C	5L096
		15/70	460	B	
		15/62	330	G	

審査請求　未請求　請求項の数12　ＯＬ　（全17頁）

(21)出願番号　特願平10－203583

(22)出願日　平成10年７月17日(1998.7.17)

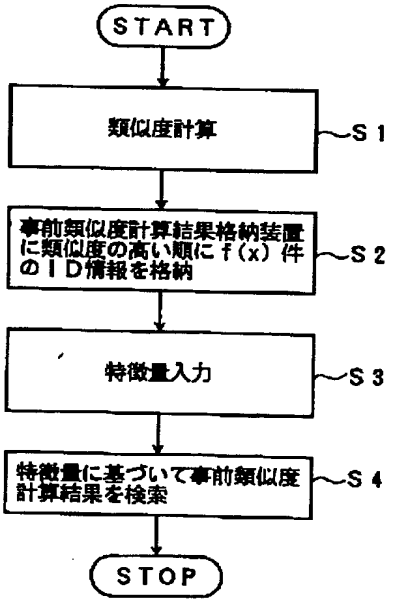
(71)出願人　000004226
日本電信電話株式会社
東京都千代田区大手町二丁目３番１号
(72)発明者　赤間　浩樹
東京都新宿区西新宿三丁目19番２号　日本電信電話株式会社内
(72)発明者　佐藤　路恵
東京都新宿区西新宿三丁目19番２号　日本電信電話株式会社内
(74)代理人　100070150
弁理士　伊東　忠彦

最終頁に続く

(54)【発明の名称】類似特徴量の検索方法及び装置及び類似特徴量の検索プログラムを格納した記憶媒体

(57)【要約】
【課題】　事前類似度計算結果がディスク上または、部分的にディスク上に存在するような複雑な構造を持っていても、さらに、データベース全体の特徴量データ件数が多くとも高速な検索が可能な類似特徴量の検索方法及び装置及び類似特徴量の検索プログラムを格納した記憶媒体を提供する。
【解決手段】　本発明は、予めデータベース内の全ての特徴量をキーとし、データベース内における類似計算を行い、他の特徴量との類似度を計算し、類似度の高い順に上位ｆ（ｘ）件分のＩＤ情報を、類似度順付で、あるいは、必要に応じて該類似度の値と共に、事前類似度計算結果格納装置に格納しておき、検索キー特徴量としてデータベース内の特徴量が与えられた場合、該検索キー特徴量に対する事前類似度計算結果を検索結果として返却する。

本発明の原理を説明するための図



1

【特許請求の範囲】

【請求項１】　マルチメディアデータに対する類似検索システムやテキストの類似検索システムに用いられる類似特徴量の検索方法において、

予めデータベース内の全ての特徴量をキーとし、

前記データベース内における類似計算を行い、他の特徴量との類似度を計算し、

前記類似度の高い順に上位ｆ（ｘ）件分のＩＤ情報を、類似度順付で、あるいは、必要に応じて該類似度の値と共に、事前類似度計算結果格納装置に格納しておき、

検索キー特徴量として前記データベース内の特徴量が与えられた場合、該検索キー特徴量に基づいて前記事前類似度計算結果格納装置を検索して、事前類似度計算結果を検索結果として返却することを特徴とする類似特徴量の検索方法。

【請求項２】　前記検索キー特徴量として前記データベース内に存在することが分からない特徴量が与えられた場合に、

前記特徴量に最も類似する前記データベース内の特徴量を最近傍検索装置により検索し、

検索結果の特徴量に基づいて前記事前類似度計算結果格納装置を検索して、事前類似度計算結果を検索結果として返却する請求項１記載の類似特徴量の検索方法。

【請求項３】　特徴量データの追加がある場合に、追加されたデータに関しては、追加特徴量データ管理装置で管理を行い、

検索キー特徴量が与えられた場合には、

前記事前類似度計算結果格納装置を検索した結果と、前記追加特徴量データ管理装置からの検索結果を類似度順にマージした結果を検索結果として返却する請求項１乃至２記載の類似特徴量の検索方法。

【請求項４】　特徴量データの追加がある場合に、前記追加特徴量データ管理装置内の特徴量データ数が特定値ｔを越えた後に、または、特定間隔の時間経過を含むタイミングにより、追加特徴量データを含めたデータベース内の全てのデータに関し、前記事前類似度計算結果の再計算を検索を行う処理とは、独立にまたは、並列に行い、

計算が完了した時点で、前記事前類似度計算結果、及び前記追加特徴量データ管理装置のデータの更新を行う請求項３記載の類似特徴量の検索方法。

【請求項５】　マルチメディアデータに対する類似検索システムやテキストの類似検索システムに用いられる類似特徴量の検索装置であって、

全ての特徴量をキーとするデータベースと、

前記データベース内における類似計算を行い、他の特徴量との類似度を計算する類似度計算手段と、

前記類似度計算手段で求められた前記類似度の高い順に上位ｆ（ｘ）件分のＩＤ情報を、類似度順付で、あるいは、必要に応じて該類似度の値と共に格納する、事前類似度計算結果格納手段と、

2

検索キー特徴量として前記データベース内の特徴量が与えられた場合、該検索キー特徴量に対する事前類似度計算結果を前記事前類似度計算結果格納手段を検索することにより取得して、検索結果として返却する事前類似度計算結果検索手段とを有することを特徴とする類似特徴量の検索装置。

【請求項６】　前記検索キー特徴量として前記データベース内に存在することが分からない特徴量が与えられた場合に、前記特徴量に最も類似する前記データベース内の特徴量を検索する最近傍検索手段を更に有し、

前記事前類似度計算結果検索手段は、前記最近傍検索手段の検索結果の特徴量に対する事前類似度計算結果を検索結果として返却する請求項５記載の類似特徴量の検索装置。

【請求項７】　特徴量データの追加がある場合に、追加されたデータに関して管理する追加特徴量データ管理手段と、

検索キー特徴量が与えられた場合には、前記事前類似度計算結果格納手段からの結果と、前記追加特徴量データ管理手段からの検索結果を類似度順にマージした結果を検索結果として返却するマージ手段を有する請求項５乃至６記載の類似特徴量の検索装置。

【請求項８】　特徴量データの追加がある場合に、前記追加特徴量データ管理手段内の特徴量データ数が特定値ｔを越えた後に、または、特定間隔の時間経過を含むタイミングにより、追加特徴量データを含めたデータベース内の全てのデータに関し、前記事前類似度計算結果検索手段とは独立または、並列に事前類似度計算を行う再計算手段と、

前記再計算手段の計算が完了した時点で、前記事前類似度計算結果格納手段、及び前記追加特徴量データ管理手段のデータの更新を行う更新手段を有する請求項７記載の類似特徴量の検索装置。

【請求項９】　マルチメディアデータに対する類似検索システムやテキストの類似検索システムに用いられる類似特徴量の検索プログラムを格納した記憶媒体であって、

データベース内の全ての特徴量をキーとするデータベース内における類似計算を行い、他の特徴量との類似度を計算する類似度計算プロセスと、

前記類似度計算プロセスで求められた前記類似度の高い順に上位ｆ（ｘ）件分のＩＤ情報を、類似度順付で、あるいは、必要に応じて該類似度の値と共に事前類似度計算結果格納手段に格納する事前類似度計算結果格納制御プロセスと、検索キー特徴量として前記データベース内の特徴量が与えられた場合、該検索キー特徴量に対する事前類似度計算結果を前記事前類似度計算結果格納手段を検索することにより取得して、検索結果として返却する事前類似度計算結果検索プロセスとを有することを特

徴とする類似特徴量の検索プログラムを格納した記憶媒体。

【請求項１０】　前記検索キー特徴量として前記データベース内に存在することが分からない特徴量が与えられた場合に、前記特徴量に最も類似する前記データベース内の特徴量を検索する最近傍検索プロセスを更に有し、前記事前類似度計算結果検索プロセスは、前記最近傍検索プロセスの検索結果の特徴量に対する事前類似度計算結果を検索結果として返却する請求項９記載の類似特徴量の検索プログラムを格納した記憶媒体。

【請求項１１】　検索キー特徴量が与えられた場合には、前記事前類似度計算結果格納手段からの結果と、特徴量データの追加がある場合に、追加されたデータに関して管理する追加特徴量データ管理手段からの検索結果を類似度順にマージした結果を検索結果として返却するマージプロセスを有する請求項８乃至１０記載の類似特徴量の検索プログラムを格納した記憶媒体。

【請求項１２】　特徴量データの追加がある場合に、前記追加特徴量データ管理手段内の特徴量データ数が特定値ｔを越えた後に、または、特定間隔の時間経過を含むタイミングにより、追加特徴量データを含めたデータベース内の全てのデータに関し、前記事前類似度計算結果検索プロセスとは独立または、並列に事前類似度計算を行う再計算プロセスと、

前記再計算プロセスの計算が完了した時点で、前記事前類似度計算結果格納手段、及び前記追加特徴量データ管理手段のデータの更新を行う更新プロセスを含む請求項１１記載の類似特徴量の検索プログラムを格納した記憶媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、類似特徴量の検索方法及び装置及び類似特徴量の検索プログラムを格納した記憶媒体に係り、特に、画像、映像、モーション、音楽、音声などのマルチメディアデータに対する類似検索システムの実現やテキストの類似検索システムに用いられる類似特徴量の検索方法及び装置及び類似特徴量の検索プログラムを格納した記憶媒体に関する。詳しくは、インターネット上の画像のように、大量で、その量が日々増加するような対象に対し、高速な類似検索を実現することに用いるための類似特徴量の検索方法及び装置及び類似特徴量の検索プログラムを格納した記憶媒体に関する。

【０００２】

【従来の技術】最初に多次元特徴量データについて説明する。画像検索、音楽検索などに代表される検索は、従来のＲＤＢＭＳが対象としてきた一致検索や範囲検索とは異なり、多次元特徴量（次元数は１以上）の類似検索である。

【０００３】ここで、一致検索とは、データベース内の列に対する検索キー値が与えられた時、それと一致する値を持つ全行、または、行ＩＤを検索結果とする検索をいう。範囲検索とは、データベース内の列に対し、検索キーとしての値と共に、検索条件として大小関係が与えられ、その関係を満足する値を持つデータベース内の全行、または、行ＩＤを検索結果とする検索をいう。

【０００４】一方、類似検索とは、１次元以上の多次元特徴量（一般に単に特徴量と呼ぶ。複数の数値からなるためベクトルと呼ぶこともある）をデータベース格納の対象とし、検索キーとして与えられた特徴量キーに対し、その特徴量間の距離等を計算することにより類似度を求め、最も類似度の高い順に上位ｆ（ｘ）件の行を求めるような検索を行う。

【０００５】特徴量としては、画像や音楽等マルチメディア情報の内容特徴などのこともあるし、地図座標のこともあるし、テキスト内のキーワードの重みのこともある。類似検索は、範囲検索の対象を１次元の値から多次元のベクトル値に拡張した場合に似ているが、その返却値の考え方が異なり、範囲検索の場合は、範囲条件が明確に指定され、その条件を満たす行は全て検索結果となるものの、類似検索の場合は、一般には、類似の高い順に上位ｆ（ｘ）件を返すという指定が用いられる。

【０００６】以下の明細書中の説明において、上位ｆ（ｘ）件と記述した場合に、それは抽象化された値を示しており、単に、利用者が指定した特定の値ｋ、システムが予め持つ特定の値ｋ、また、最大ｋや最小ｋ、データベース内の全データ数、利用者、システムまたは、データベースの状態から得られる情報等から計算によって求められた利用者または、システムまたは、データベースの状態から得られる情報等から計算によって求められた値のように、別の手段で計算される閾値ｋでもよい。また、図等で１つのフローチャート内に複数のｆ（ｘ）という表記があっても、それらは独立な値を持ってもよい。

【０００７】図１０は、類似検索の例を説明するための図である。この例の特徴量は２次元で、データベース内には６件の特徴量データが登録されている。この利用者から与えられた検索キー特徴量（０．５、０．６）を入力した場合、データベース内の各特徴量とのユークリッド距離を計算し、その距離の近い順に並べ替え、その中の上位何件かを検索結果として利用者に返却する。

【０００８】次に、高速化について説明する。最も単純な類似検索では、検索キー特徴量とデータベース内の全特徴量データとの類似度計算が検索実行時に行われる。ところで、この特徴量が１次元の場合には、従来のＲデータベースＭＳの範囲検索で利用されていたような手法（Ｂ＋Ｔｒｅｅ等）を用いることで高速検索が可能になる。

【０００９】しかし、類似検索では、一般には特徴量は２以上の次元数となるため、上記の手法は利用できな

い。そこで、その高速化のための索引手法には以下のような手法が用いられる。図１１、図１２は、Ｒ－ｔｒｅｅの例を示しており、図１１は、従来のＲ－ｔｒｅｅの特徴量空間分割を説明するための図であり、図１２は、従来のＴ－ｔｒｅｅの木構造を説明するための図である。構成される木の各ノードは、どの次元で分割したかという情報と、その範囲の情報を持つ。各分割は、その中に含まれる特徴量点の個数が同程度になるように調整されている。図１３、図１４は、ＰＲ－ｑｕａｄｔｒｅｅの例を示しており、図１３は、従来のＰＲ－ｑｕａｄｔｒｅｅの特徴量空間分割を説明するための図であり、図１４は、従来のＰＲ－ｑｕａｄｔｒｅｅの木構造を説明するための図である。空間は常にＸ－Ｙ平面で４つに分割され、分割後の領域に指定個数以上の特徴量点が存在する場合は、さらに４分割が行われていく。

【００１０】それぞれに関し、各種の改良バージョンが提案されているが（参考：Volker Gaedo, Multidimensional Access Methods ）、一般には、大量のデータに対しても、その木を平衡状態を維持するＲ－ｔｒｅｅ、及びその改良系が高速性、汎用性に優れている。本発明では、これらの多次元空間を分割し、木状にした構造をもつ索引を木状索引と呼ぶことにする。

【００１１】図１５は、従来の木状索引を使った類似検索のフローチャートであり、木状索引の構築時の流れ、及び木状索引を使った類似検索時の流れを示している。索引構築時は、特徴量の部分集合をデータベース全体の特徴量とし（ステップ１０）、特徴量数または、リンク数が（木のノード内数）以上であれば、特徴量の部分集合に対して以下の処理を行う（ステップ１１）、分割基準を決定し（ステップ１２）、分割基準によって特徴量集合をｎ個の分割し（ステップ１３）、個々の集合に対し、再帰的に繰り返す（ステップ１４）。再帰終了の場合には階層的分類結果を索引として登録する（ステップ１５）。

【００１２】検索実行時は、検索キー特徴量を入力し（ステップ２０）、与えられた特徴量がどの分類に相当するか、分類基準に従って索引を辿る（ステップ２１）。

【００１３】

【発明が解決しようとする課題】しかしながら、上記従来の木状索引による高速な検索手法には以下のような問題がある。最初に高次元数特徴量データでの検索速度における観点から説明する。従来の木状検索手法は、特徴量データ数の増加に対して、その検索速度の増加を押さえることを主な目的としている。つまり、特徴量データを木構造で管理することで、特徴量データ間の比較階数をｌｏｇのオーダとし、特徴量データ数の増加に対する検索性能を維持する。

【００１４】しかし、これら従来の木状索引構造は、次元数の増加に対しては考慮されておらず、例えば、Ｒ－ｔｒｅｅ等では、数次元程度で最も威力を発揮し、２０次元を越えるとその性能はかなり悪くなることが知られている。これは、地理情報等、低次元の応用には充分だが、マルチメディア情報等の高次元の応用には不十分である。

【００１５】次に、高度類似基準への対処における観点から説明する。従来の索引手法は、マンハッタン距離（市街地距離）やユークリッド距離のように数学的に距離の公理を満たす単純な類似度基準を想定している。これらの類似度基準により、事前にデータベース内のデータ間の関係を各次元軸をもとにクラスタリングした場合には、そのクラスタリング結果空間の中で近いデータ同士は、その元となる類似度基準でも近いという性質があり、事前に木状索引の作成が可能になる。しかし、その類似度基準が与えられた検索キーデータに依存し、各次元を元に事前にクラスタリングすることが意味をなさない場合、例えば、ヒストグラム、インターセクション（参考：Maichael J. Swain, Indexing Via Color Histogram ）や、非対象類似度（参考：赤間、オブジェクトの類似度算出方法及び類似オブジェクト検索装置、特願平９－０６０９９９）といった、マルチメディア情報の特徴量に合った複雑な類似度基準には対応できない。なお、本明細書では、距離をより一般化した用語として類似度を用いている。

【００１６】次に、近傍順検索時の検査速度の観点から説明する。木状に構成された索引では、最近傍データを検出するのは容易である。しかし、一般的な類似検索においては、最も類似するものだけを検索するに留まらず、似ている順に上位ｆ（ｘ）件の結果を求めることが多い。その場合、木状に管理されたデータにおいては、木の枝や葉を順に辿り、候補の中のデータに関して、再度、類似度の計算を行う必要がある。また、これは、特徴量データ数が増加し、データがメモリ上ではなく、ディスク上にある場合は、かなりの速度低下要因となる。

【００１７】最後に、実装法の観点から説明する。木の平衡状態を維持する等、アルゴリズムが複雑で実装が困難である。本発明は、上記の点に鑑みなされたもので、事前類似度計算結果がディスク上または、部分的にディスク上に存在するような複雑な構造を持っていても、さらに、データベース全体の特徴量データ件数が多くとも高速な検索が可能な類似特徴量の検索方法及び装置及び類似特徴量の検索プログラムを格納した記憶媒体を提供することを目的とする。

【００１８】

【課題を解決するための手段】図１は、本発明の原理を説明するための図である。本発明（請求項１）は、マルチメディアデータに対する類似検索システムやテキストの類似検索システムに用いられる類似特徴量の検索方法において、予めデータベース内の全ての特徴量をキーとし、データベース内における類似計算を行い、他の特徴

量との類似度を計算し（ステップ１）、類似度の高い順に上位ｆ（ｘ）件分のＩＤ情報を、類似度順付で、あるいは、必要に応じて該類似度の値と共に、事前類似度計算結果格納装置に格納しておき（ステップ２）、検索キー特徴量としてデータベース内の特徴量が与えられた場合（ステップ３）、該検索キー特徴量に基づいて事前類似度計算結果格納装置を検索して、事前類似度計算結果を検索結果として返却する（ステップ４）。

【００１９】本発明（請求項２）は、検索キー特徴量としてデータベース内に存在することが分からない特徴量が与えられた場合に、特徴量に最も類似するデータベース内の特徴量を最近傍検索装置により検索し、検索結果の特徴量に基づいて事前類似度計算結果格納装置を検索して、事前類似度計算結果を検索結果として返却する。

【００２０】本発明（請求項３）は、特徴量データの追加がある場合に、追加されたデータに関しては、追加特徴量データ管理装置で管理を行い、検索キー特徴量が与えられた場合には、事前類似度計算結果格納装置を検索した結果と、追加特徴量データ管理装置からの検索結果を類似度順にマージした結果を検索結果として返却する。

【００２１】本発明（請求項４）は、特徴量データの追加がある場合に、追加特徴量データ管理装置内の特徴量データ数が特定値ｔを越えた後に、または、特定間隔の時間経過を含むタイミングにより、追加特徴量データを含めたデータベース内の全てのデータに関し、事前類似度計算結果の再計算を検索を行う処理とは、独立にまたは、並列に行い、計算が完了した時点で、事前類似度計算結果、及び追加特徴量データ管理装置のデータの更新を行う。

【００２２】図２は、本発明の原理構成図である。本発明（請求項５）は、マルチメディアデータに対する類似検索システムやテキストの類似検索システムに用いられる類似特徴量の検索装置であって、全ての特徴量をキーとするデータベース１０と、データベース１０内における類似計算を行い、他の特徴量との類似度を計算する類似度計算手段２０と、類似度計算手段２０で求められた類似度の高い順に上位ｆ（ｘ）件分のＩＤ情報を、類似度順付で、あるいは、必要に応じて該類似度の値と共に格納する、事前類似度計算結果格納手段３０と、検索キー特徴量としてデータベース１０内の特徴量が与えられた場合、該検索キー特徴量に対する事前類似度計算結果を事前類似度計算結果格納手段３０を検索することにより取得して、検索結果として返却する事前類似度計算結果検索手段４０とを有する。

【００２３】本発明（請求項６）は、検索キー特徴量としてデータベース１０内に存在することが分からない特徴量が与えられた場合に、特徴量に最も類似するデータベース１０内の特徴量を検索する最近傍検索手段を更に有し、事前類似度計算結果検索手段４０は、最近傍検索手段の検索結果の特徴量に対する事前類似度計算結果を検索結果として返却する。

【００２４】本発明（請求項７）は、特徴量データの追加がある場合に、追加されたデータに関して管理する追加特徴量データ管理手段と、検索キー特徴量が与えられた場合には、事前類似度計算結果格納手段３０からの結果と、追加特徴量データ管理手段からの検索結果を類似度順にマージした結果を検索結果として返却するマージ手段を有する。

【００２５】本発明（請求項８）は、特徴量データの追加がある場合に、追加特徴量データ管理手段内の特徴量データ数が特定値ｔを越えた後に、または、特定間隔の時間経過を含むタイミングにより、追加特徴量データを含めたデータベース１０内の全てのデータに関し、事前類似度計算結果検索手段４０とは独立または、並列に事前類似度計算を行う再計算手段と、再計算手段の計算が完了した時点で、事前類似度計算結果格納手段３０、及び追加特徴量データ管理手段のデータの更新を行う更新手段を有する。

【００２６】本発明（請求項９）は、マルチメディアデータに対する類似検索システムやテキストの類似検索システムに用いられる類似特徴量の検索プログラムを格納した記憶媒体であって、データベース内の全ての特徴量をキーとするデータベース内における類似計算を行い、他の特徴量との類似度を計算する類似度計算プロセスと、類似度計算プロセスで求められた類似度の高い順に上位ｆ（ｘ）件分のＩＤ情報を、類似度順付で、あるいは、必要に応じて該類似度の値と共に事前類似度計算結果格納手段に格納する事前類似度計算結果格納制御プロセスと、検索キー特徴量としてデータベース内の特徴量が与えられた場合、該検索キー特徴量に対する事前類似度計算結果を事前類似度計算結果格納手段を検索することにより取得して、検索結果として返却する事前類似度計算結果検索プロセスとを有する。

【００２７】本発明（請求項１０）は、検索キー特徴量としてデータベース内に存在することが分からない特徴量が与えられた場合に、特徴量に最も類似するデータベース内の特徴量を検索する最近傍検索プロセスを更に有し、事前類似度計算結果検索プロセスは、最近傍検索プロセスの検索結果の特徴量に対する事前類似度計算結果を検索結果として返却する。

【００２８】本発明（請求項１１）は、検索キー特徴量が与えられた場合には、事前類似度計算結果格納手段からの結果と、特徴量データの追加がある場合に、追加されたデータに関して管理する追加特徴量データ管理手段からの検索結果を類似度順にマージした結果を検索結果として返却するマージプロセスを有する。本発明（請求項１２）は、特徴量データの追加がある場合に、追加特徴量データ管理手段内の特徴量データ数が特定値ｔを越えた後に、または、特定間隔の時間経過を含むタイミン

グにより、追加特徴量データを含めたデータベース内の全てのデータに関し、事前類似度計算結果検索プロセスとは独立または、並列に事前類似度計算を行う再計算プロセスと、再計算プロセスの計算が完了した時点で、事前類似度計算結果格納手段、及び追加特徴量データ管理手段のデータの更新を行う更新プロセスを含む。

【００２９】上記のように、本発明は、類似度の高い順に上位ｆ（ｘ）件分のＩＤ情報に類似度順が付与された事前類似度計算結果に対する最近傍検索処理は、既にデータベース内に存在する値に対する一致検索となるため、その索引方法としては、従来のＢ－Ｔｒｅｅ、Ｂ＋Ｔｒｅｅ、ハッシュ等のごく一般的な（容易な）手法を利用することで実現できる。また、近傍順検索については、事前に計算してある結果をそのまま返却するだけの処理となるため、その結果が例え、ディスク上にあろうが、部分的にディスク上に存在するような複雑な構造を持っていようが、非常に高速に検索が可能になる。

【００３０】また、検索実行時に次元数に依存する類似度計算を行うことがないため、次元数の増加に対しても性能が劣化することが少なく、高速である。さらに、索引の構造の中に距離に依存した部分がないので、特殊な類似性基準にも対応できる。また、データベース内の特徴量のみを対象とする場合には、最も類似する特徴量は、必ず自分自身であるため、一般には出力するか否かについてシステムに依存するが、データベース外特徴量を対象とする場合には、通常、最も類似する特徴量を出力する必要がある。

【００３１】また、特徴データの追加がある場合でも事前類似度計算結果格納手段からの結果と追加特徴量データ管理手段に格納されている検索結果を類似度順にマージして、上位ｆ（ｘ）件を検索結果として出力することができる。これにより、追加データのあるシステムの場合においても事前類似度計算結果を索引として利用することが可能となる。

【００３２】

【発明の実施の形態】以下の説明において、特徴量データをデータベース内に存在する特徴量（これをデータベース内特徴量と呼ぶ）と、データベース内に存在しない特徴量（これをデータベース外特徴量と呼ぶ）の２種類に分けて考える。例えば、類似画像検索システムにおいて、データベース外特徴量を検索キーとして利用する例としては、スケッチ入力画像を検索キーとする場合や、ディジタルカメラ画像を検索キーとする場合がある。

【００３３】一方、データベース内部特徴量のＩＤを検索キーとして利用する例としては、システムが利用者にランダムに提示した画像を検索キーとする場合や、キーワード検索等の他の手法で検索した画像を検索キーとする場合や、一度検索した結果を利用してナビゲーション的に繰り返し検索する場合などがある。類似検索の索引の処理を、検索キー特徴量に最も類似する特徴量を求める処理である最近傍検索と、最近傍検索で求めた特徴量から順に近い特徴量を求めていく近傍順検索の２つの処理を分けて考えると、データベース外特徴量を検索キーとする類似検索では、最近傍検索と近傍順検索の両方が必要であり、データベース内特徴量のＩＤを検索キーとする類似検索では、近傍順検索のみ必要である。

【００３４】なお、検索キーとしてデータベース内特徴量そのものが与えられた場合においても、一致検索によってデータベース内特徴量ＩＤに変換することが可能であるため、最近傍検索は必要ない。本発明では、主に近傍順検索の処理部分の高速化を対象とする。図３は、本発明の類似特徴量検索装置の構成を示す。

【００３５】同図に示す類似特徴量検索装置は、データベース１０、類似度計算部２０、事前類似度計算結果格納部３０、検索部４０、検索キー入力部５０、特徴量種別判定部５５、出力部６０、最近傍検索部７０、追加特徴量データ管理部８０、マージ部９０から構成される。データベース１０は、全ての特徴量をキーとして、ＩＤ情報及びデータと共に格納する。

【００３６】類似度計算部２０は、データベース１０内における類似計算を行い、他の特徴量との類似度を計算し、類似度の高い順に上位ｆ（ｘ）件分のＩＤ情報に類似度順を付与してデータベース１０に事前類似度計算結果格納部３０に格納する。必要によっては、当該類似度の値と共に、事前類似度計算結果格納部３０に格納する。

【００３７】事前類似度計算結果格納部３０は、類似度計算部２０により求められた類似度計算結果（類似度順、類似度が付与されたＩＤ情報）を格納する。検索部４０は、検索キー入力部５０により与えられた検索キー特徴量として特徴量が与えられると、事前類似度計算結果格納部３０を検索して、上位ｆ（ｘ）件を検索結果として出力部６０より出力する。

【００３８】検索キー入力部５０は、検索キー特徴量として特徴量を入力する。特徴量種別判定部５５は、検索キー入力部５０から入力された特徴量がデータベース１０にあるか、データベース１０外にあるかを判定する。出力部６０は、検索部４０、最近傍検索部７０及びマージ部８０で求められた検索結果を出力する。

【００３９】最近傍検索部７０は、検索キー特徴量として検索キー入力部５０からデータベース１０内に存在するか否かが分からない特徴量が与えられた場合には、それに最も類似するデータベース内の特徴量をＲ－ｔｒｅｅ等を用いて検索し、その結果の特徴量に対する事前類似度計算結果格納部３０から検索して、上位ｆ（ｘ）件を検索結果として返却する。

【００４０】追加特徴量データ管理部８０は、検索キー入力部５０から入力された特徴量データを格納する。マージ部９０は、検索部４０が事前類似度計算結果格納部３０から検索した検索結果と、追加特徴量データ管理部

８０から検索した検索結果とをマージする。次に、上記の構成における動作を説明する。

【００４１】図４は、本発明の検索構築時及び検索実行時の動作を示すフローチャートである。まず、最初に検索構築時の動作について説明する。

ステップ１０１）　データベース１０内における全特徴量に対して以下の処理を繰り返す。

【００４２】ステップ１０２）　類似度計算部２０は、データベース１０内における類似度計算を行い、他の特徴量との類似度の計算を行い、類似度の高い順に上位ｆ（ｘ）件分のＩＤ情報を、類似度順、類似度の値を求める。

ステップ１０３）　類似度計算部２０により求められた結果を、特徴量または、そのＩＤをキーとして事前類似度計算結果格納部３０に格納する。

【００４３】次に、検索実行時の動作について説明する。

ステップ２０１）　データベース１０内の特徴量を検索キー特徴量として検索キー入力部５０より入力される。

ステップ２０２）　検索部４０は、入力された特徴量または、そのＩＤをキーとして確定検索方式により事前類似度計算結果格納部３０に対して検索を行い、検索結果を取得する。

【００４４】ステップ２０３）　検索部４０は、上位ｆ（ｘ）件分の結果を出力部６０に出力する。このときの事前類似度計算結果に対する最近傍検索の処理は、すでにデータベース１０内に存在する値に対する一致検索となるため、その検索方法としては、既存のＢ－Ｔｒｅｅ、Ｂ＋Ｔｒｅｅ、ハッシュ等の一般的な（容易な）手法を利用することで実現できる。

【００４５】また、近傍順検索については、事前に計算してある結果をそのまま返却するのみの処理となるため、その結果が例え、データベース１０上にあろうが、部分的にデータベース１０上に存在するような複雑な構造を持っていようが、非常に高速に検索が可能となる。また、検索実行時に次元数に依存する類似度計算を行うことがないため、次元数の増加に対しても性能が劣化することが少なく、高速である。

【００４６】さらに、索引の構造の中に距離に依存した部分がないので、特殊な類似性基準にも対応できる。次に、最近傍検索の処理について説明する。データベース１０内特徴量のみを対象とする場合には、最も類似する特徴量は必ずデータベース１０内にあるため、一般には出力するか否かについてシステムに依存するが、データベース外特徴量を対象とする場合には、通常最も類似する特徴量を出力する必要がある。

【００４７】図５は、本発明の最近傍検索の処理を示すフローチャートである。

ステップ３０１）　検索キー入力部５０から検索キー特徴量として、データベース１０内に存在することが分からない特徴量が与えられる。

ステップ３０２）　特徴量種別判定部５５において、入力された特徴量がデータベース１０にあるか、データベース１０外にあるかを判定し、データベース１０内にある場合にはステップ３０３に移行し、データベース１０外にある場合にはステップ３０４に移行する。

【００４８】ステップ３０３）　入力された特徴量がデータベース１０外にある場合には、最近傍検索部７０において、与えられた特徴がどの分類に相当するかを分類基準にしたがって検索を辿り、最も近い特徴量のＩＤを取得し、ステップ３０４に移行する。

ステップ３０４）　入力された特徴量がデータベース１０内にある場合には、検索部４０は、入力された特徴量または、最近傍検索部７０により求められた特徴量のＩＤをキーとして確定検索方式により事前類似度計算結果格納部３０に対して検索を行い、結果を出力部６０に出力し、ステップ３０５に移行する。

【００４９】ステップ３０５）　出力部６０において、上位ｆ（ｘ）件分の結果を出力する。次に、特徴量データの追加がある場合に対処する処理を説明する。図６は、本発明の特徴量データの追加がある場合の処理を示すフローチャートである。

【００５０】ステップ４０１）　まず、特徴量データの追加時の処理として、特徴量のデータの追加がある場合には、特徴量データの追加と索引の再構成を行い、追加特徴量データ管理部８０に格納する。

ステップ５０１）　検索実行時の処理として、検索キー特徴量が検索キー入力部５０から入力される。

【００５１】ステップ５０２）　検索部４０は、検索キー特徴量を用いて、事前類似計算結果格納部３０から上位ｆ（ｘ）件の類似検索を行う。

ステップ５０３）　さらに、検索部４０は、追加特徴量データ管理部８０から上位ｆ（ｘ）件以内の類似検索を行う。

ステップ５０４）　マージ部９０は、ステップ５０２とステップ５０３で求められた検索結果を距離順に整列させる。

【００５２】ステップ５０５）　出力部６０から上位ｆ（ｘ）件分の結果を返却する。また、特徴量データの追加がある場合において、追加特徴量データ管理部８０内の特徴量データ数が特定値ｔを越えた後に、または、特定間隔の時間経過等のタイミングにより、その追加特徴量データを含むデータベース１０内のデータに関し、検索部４０における事前類似度計算結果格納部３０による検索処理とは独立して、計算が完了した時点で、事前類似度計算結果格納部３０及び追加特徴量データ管理部８０のデータの更新を行う。

【００５３】これにより、追加データのあるシステムの場合でも、事前類似度計算結果を索引として利用することが可能となる。但し、高度な類似度を用いる場合や高

次元特徴量を扱う場合や、データ数が少ない場合には、Ｒ－ｔｒｅｅ等の木状索引を利用せずに、全件処理を行う方が望ましい。

【００５４】なお、本明細書では、特徴量種が１種類の場合を想定して記述しているが、２種以上の特徴量が存在し、それらを独立に検索するような場合にも、複数の事前類似度計算結果を持つことで同様に適用できる。特徴量種別としては、画像の場合、色相、彩度、輝度、テクスチャ、大きさ等、画像オブジェクトの場合には、さらに、形、位置、傾き等多種存在する。

【００５５】また、本明細書では、１種類の特徴量に対し、１種類の類似度基準を前提として記述しているが、複数の類似度基準（または、距離基準）を切り替えて検索を可能にするシステムに対応するため、事前類似度計算結果を類似度基準の種類数分だけ用意すればよい。

【００５６】

【実施例】以下、図面と共に本発明の実施例を説明する。

［第１の実施例］図７は、本発明の第１の実施例の事前類似度計算結果の例を説明するための図であり、図８は、本発明の第１の実施例の総当たりによる事前類似度計算の例を説明するための図である。

【００５７】図７、図８を用いて事前類似度計算結果の作成方法、及びその検索時の利用方法について説明する。まず、事前類似度計算結果を作成するため、データベース１０内の全特徴量に対して以下の処理を繰り返す。始めに、キーをＩＤ１の（０．３、０．３）とし、ＩＤ２～ＩＤ６までのデータを対象とした類似検索を行う。その結果が図８に示されている。この例では、最も簡単な実装の場合を想定し、「ｆ（ｘ）件」として全件（この場合６件）だったとした場合で、類似検索方法は、全ての組み合わせで類似度を計算した場合とし、類似度の高い順に並んだ６つのＩＤの結果、

ＩＤ１，ＩＤ２，ＩＤ４，ＩＤ５，ＩＤ３，ＩＤ６

を求め、図７のＩＤ１の事前類似度計算結果として登録している。

【００５８】同様に、キーＩＤ２～キーＩＤ６までの処理を行った場合を図７に示す。なお、類似検索方法は、Ｒ－Ｔｒｅｅのような他の既存の類似検索用索引手法を用いた方法であっても構わない。また、事前類似度計算結果中には、必要に応じて、類似度等の情報を持つこともある。例えば、第３の実施例において後述するように、再度、類似度計算が必要な場合には、事前類似度計算結果として類似度情報を持つと効率がよい。

【００５９】次に、事前類似度計算結果を使った検索の例を示す。検索キーとして与えられた特徴量がデータベース１０内特徴量と分かる場合、通常、その情報はＩＤとして与えられ、ＩＤを使って、事前類似度計算結果から、ＩＤに割り当てられている事前類似度計算結果を得ることができる。しかし、もし、この段階で特徴量しか与えられなかった場合でも、特徴量に対し、普通のＢ－ｔｒｅｅ索引等が付与してあれば、単なる一致検索として、高速にそのＩＤを求めることができる。

【００６０】なお、事前類似度計算結果に登録してあるＩＤの件数が、検索として要求され、検索された件数より少ない場合には、本発明では、上位ｆ（ｘ）件までの部分にしか機能せず、ｆ（ｘ）件の部分については、従来手法による類似順検索が必要になる。しかし、通常は、データベース作成時にアプリケーションとして利用する最大件数が決定できるため、それを越える個数のＩＤを事前類似度計算結果に用意しておけば問題ない。

【００６１】［第２の実施例］本実施例では、与えられる検索キーが内部データベース特徴量と判断できない場合の例を示す。与えられた検索キーがデータベース内部にある特徴量と判断できない場合には、その特徴量データによる最近傍検索だけをＲ－Ｔｒｅｅのような他の従来手法を利用し、その後の近傍順検索については、本発明を利用する。

【００６２】これは、例えば、図１０に示す類似検索のように、検索キーとして（０．５，０．６）が与えられた場合、その最近傍特徴量の（０．５，０．５）を求めるまでは、従来手法を用い、その後、（０．５，０．５）の近傍順検索では、そのＩＤに登録されている事前類似度計算結果の

ＩＤ４，ＩＤ３，ＩＤ１，ＩＤ５，ＩＤ２，ＩＤ６

を検索結果とする。

【００６３】なお、厳密な類似度順の結果を得たい場合には、再度、類似度計算を行い、整列をし直すものとする。

［第３の実施例］本実施例では、特徴量データに追加が有る場合の処理を図７及び図９を用いて説明する。

【００６４】図９は、本発明の第３の実施例の追加特徴量の管理とその類似検索の例を説明するための図である。本実施例において、事前類似度計算結果は図７に示すものとし、後に追加されたデータは、図９のように追加特徴量データ管理部８０に格納される。この追加特徴量データ管理部８０には、一般には従来の木状索引等が付与され、高速化される。また、データベース外特徴量を扱う場合の最近傍検索用索引と統合されることもある。

【００６５】検索キー特徴量が与えられた場合には、事前類似度計算結果から上位ｆ（ｘ）件の類似検索結果を得、同時に、追加特徴量データ管理部８０からも最大で上位ｆ（ｘ）件の類似検索結果を得る。このとき、それらの結果に類似度情報も付与しておき、その類似度で上位ｆ（ｘ）件の類似度データを作成し、それを類似検索結果とする。

【００６６】なお、この検索キー特徴量がデータベースない特徴量の場合には、事前類似度計算結果から得られた上位ｆ（ｘ）件に対し、事前に計算された類似度を利

用することができるが、データベース外特徴量の場合には、類似度に関し、再計算が必要となる。例えば、検索キー特徴量が（０．５、０．６）の場合、第２の実施例で示したように、事前類似度計算結果を使った検索結果は、

ＩＤ４，ＩＤ３，ＩＤ１，ＩＤ５，ＩＤ２，ＩＤ６

となり、その距離の再計算を行うと、

ＩＤ４，ＩＤ３，ＩＤ１，ＩＤ５，ＩＤ２，ＩＤ６

になる。

【００６７】また、図９の追加特徴量データ管理部８０から検索した結果は、

ＩＤ１，ＩＤ２，ＩＤ３

となり、これらを類似度順にマージすると、

内ＩＤ４，内ＩＤ３，追ＩＤ１，内ＩＤ１，内ＩＤ６，内ＩＤ５，追ＩＤ２，追ＩＤ３，内ＩＤ２

となる。

【００６８】但し、事前類似度計算結果内のＩＤは、「内ＩＤ」と、追加特徴量データ管理部８０内のＩＤは、「追ＩＤ」と記載し、区別した。よって、この内の上位ｆ（ｘ）件を検索結果とすればよい。また、本発明は、上記の実施例に限定されることなく、図３に示す構成要件をプログラムとして構築し、類似特徴量検索装置として利用されるコンピュータに接続されるディスク装置や、フロッピーディスク、ＣＤ－ＲＯＭ等の可搬記憶媒体に格納しておき、本発明を実施する際に、インストールすることにより容易に本発明を実現できる。

【００６９】なお、本発明は、上記の実施例に限定されることなく、特許請求の範囲内で種々変更・応用が可能である。

【００７０】

【発明の効果】上述のように、本発明によれば、事前類似度計算結果情報が、たとえ、ディスク上にあろうが、部分的にディスク上に存在するような複雑な構造を持っていようが、データベース全体の特徴量データ件数が多かろうが、非常に高速に検索ができる。

【００７１】また、検索実行時に次元数に依存する類似度計算を行うことがないため、次元数の増加に対しても性能が劣化することが少なく、高速である。さらに、索引の構造の中に距離に依存した部分がないので、特殊な類似性基準にも対応できる。また、本発明は、近傍検索と組み合わせた、高速な類似検索が可能となる。

【００７２】さらに、本発明は、追加の特徴量データが存在する場合にも、システムの構成が可能となる。

【図面の簡単な説明】

【図１】本発明の原理を説明するための図である。

【図２】本発明の原理構成図である。

【図３】本発明の類似特徴量検索装置の構成図である。

【図４】本発明の索引構築時及び検索実行時の動作を示すフローチャートである。

【図５】本発明の最近傍検索の処理を示すフローチャートである。

【図６】本発明の特徴量データの追加がある場合の処理を示すフローチャートである。

【図７】本発明の第１の実施例の事前類似度計算結果の例を説明するための図である。

【図８】本発明の第１の実施例の総当たりによる事前類似度計算の例を説明するための図である。

【図９】本発明の第３の実施例の追加特徴量の管理とその類似検索の例を説明するための図である。

【図１０】類似検索を説明するための図である。

【図１１】従来のＲ－ｔｒｅｅの特徴量空間分割を説明するための図である。

【図１２】従来のＲ－ｔｒｅｅの木構造を説明するための図である。

【図１３】従来のＰＲ－ｑｕａｄｔｒｅｅの特徴量空間分割を説明するための図である。

【図１４】従来のＰＲ－ｑｕａｄｔｒｅｅの木構造を説明するための図である。

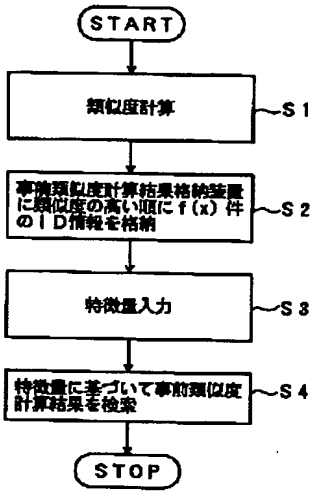
【図１５】従来の木状索引を使った類似検索のフローチャートである。

【符号の説明】

１０　データベース
２０　類似度計算手段、類似度計算部
３０　事前類似度計算結果格納手段、事前類似度計算結果格納部
４０　事前類似度計算結果検索手段、検索部
５０　検索キー入力部
５５　特徴量種別判定部
６０　出力部
７０　最近傍検索部
８０　追加特徴量データ管理部
９０　マージ部

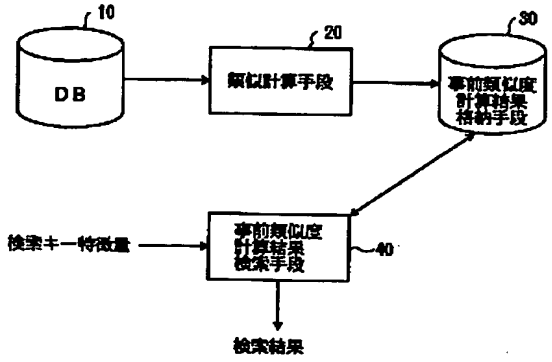
【図１】

本発明の原理を説明するための図



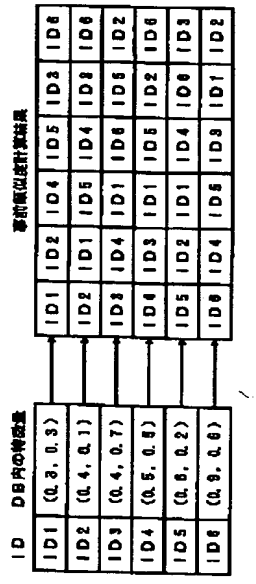
【図２】

本発明の原理構成図



【図７】

本発明の第１の実施例の事前類似度計算結果の例を説明するための図



【図８】

本発明の第１の実施例の総当たりによる事前類似度計算の例を説明するための図

keyID＝1, (0.3, 0.3)

ID=1, (0.3, 0.3), dist=0.000
ID=2, (0.4, 0.1), dist=0.224
ID=3, (0.4, 0.7), dist=0.412
ID=4, (0.5, 0.5), dist=0.283
ID=5, (0.6, 0.2), dist=0.316
ID=6, (0.9, 0.6), dist=0.571

keyID＝4, (0.5, 0.5)

ID=1, (0.3, 0.3), dist=0.283
ID=2, (0.4, 0.1), dist=0.412
ID=3, (0.4, 0.7), dist=0.224
ID=4, (0.5, 0.5), dist=0.000
ID=5, (0.6, 0.2), dist=0.316
ID=6, (0.9, 0.6), dist=0.412

keyID＝2, (0.4, 0.1)

ID=1, (0.3, 0.3), dist=0.224
ID=2, (0.4, 0.1), dist=0.000
ID=3, (0.4, 0.7), dist=0.600
ID=4, (0.5, 0.5), dist=0.412
ID=5, (0.6, 0.2), dist=0.224
ID=6, (0.9, 0.6), dist=0.707

keyID＝5, (0.6, 0.2)

ID=1, (0.3, 0.3), dist=0.316
ID=2, (0.4, 0.1), dist=0.224
ID=3, (0.4, 0.7), dist=0.539
ID=4, (0.5, 0.5), dist=0.316
ID=5, (0.6, 0.2), dist=0.000
ID=6, (0.9, 0.6), dist=0.500

keyID＝3, (0.4, 0.7)

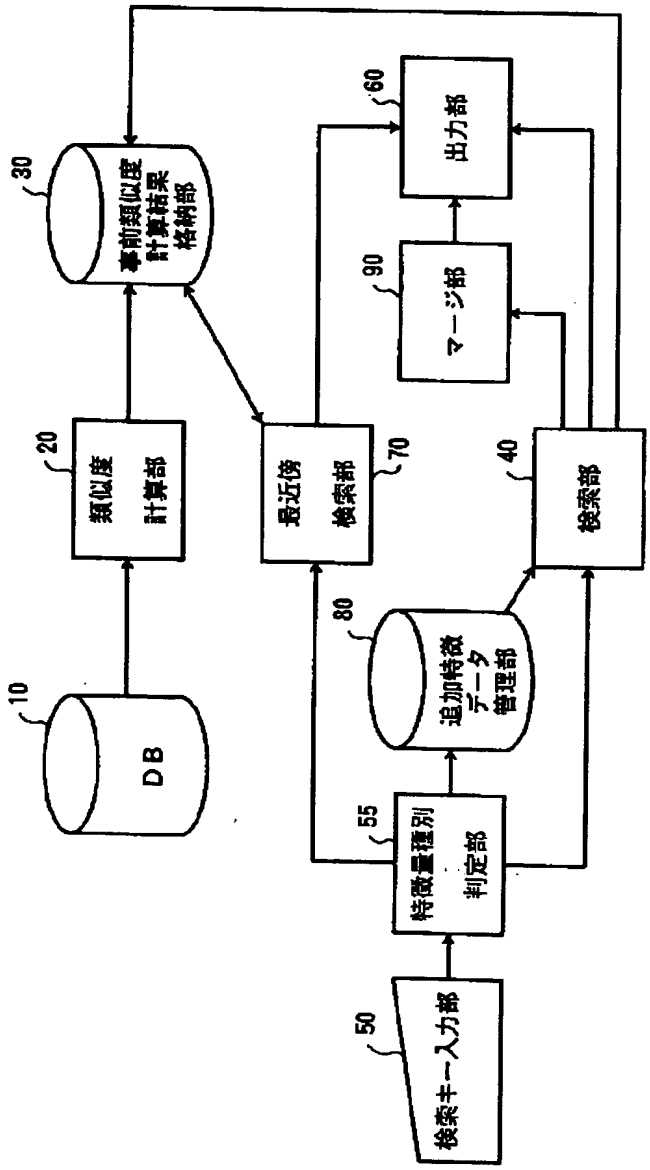
ID=1, (0.3, 0.3), dist=0.412
ID=2, (0.4, 0.1), dist=0.600
ID=3, (0.4, 0.7), dist=0.000
ID=4, (0.5, 0.5), dist=0.224
ID=5, (0.6, 0.2), dist=0.539
ID=6, (0.9, 0.6), dist=0.510

keyID＝6, (0.9, 0.6)

ID=1, (0.3, 0.3), dist=0.671
ID=2, (0.4, 0.1), dist=0.707
ID=3, (0.4, 0.7), dist=0.510
ID=4, (0.5, 0.5), dist=0.412
ID=5, (0.6, 0.2), dist=0.500
ID=6, (0.9, 0.6), dist=0.000

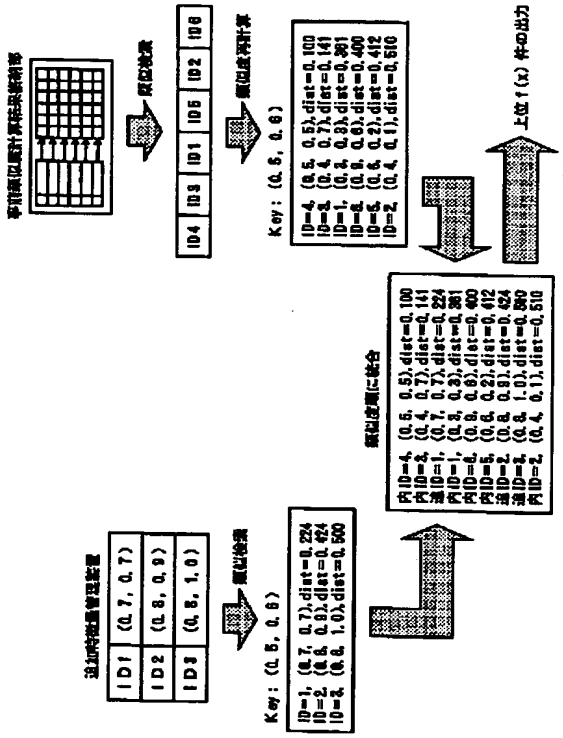
【図３】

本発明の類似特徴量検索装置の構成図



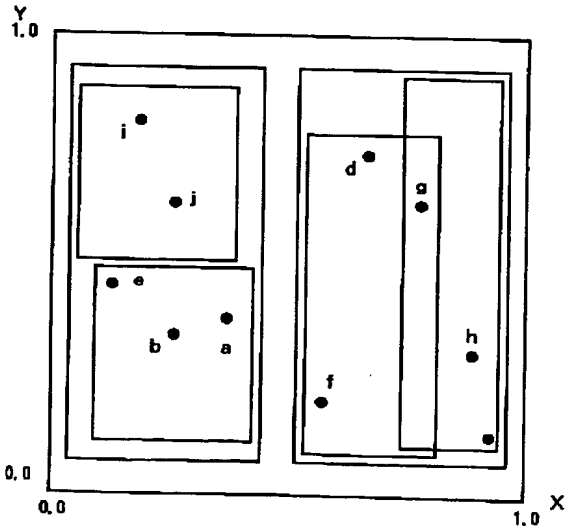
【図９】

本発明の第３の実施例の追加特徴量の管理とその類似検索の例を説明するための図



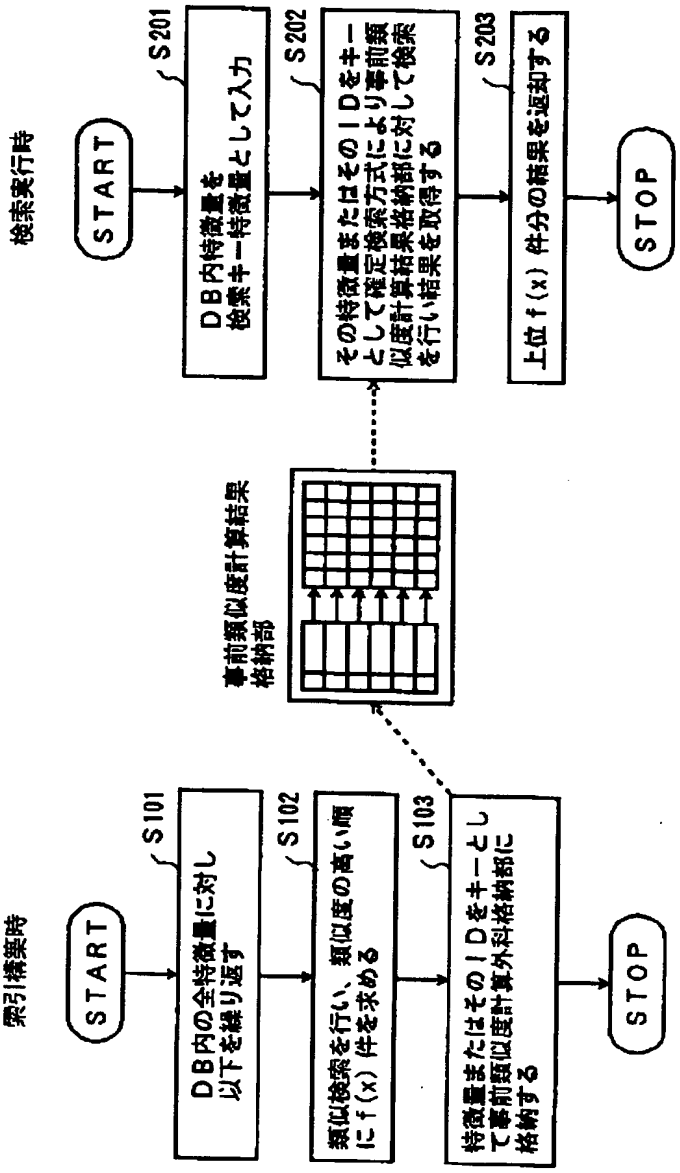
【図１１】

従来のＲ－ｔｒｅｅの特徴量空間分割を説明するための図



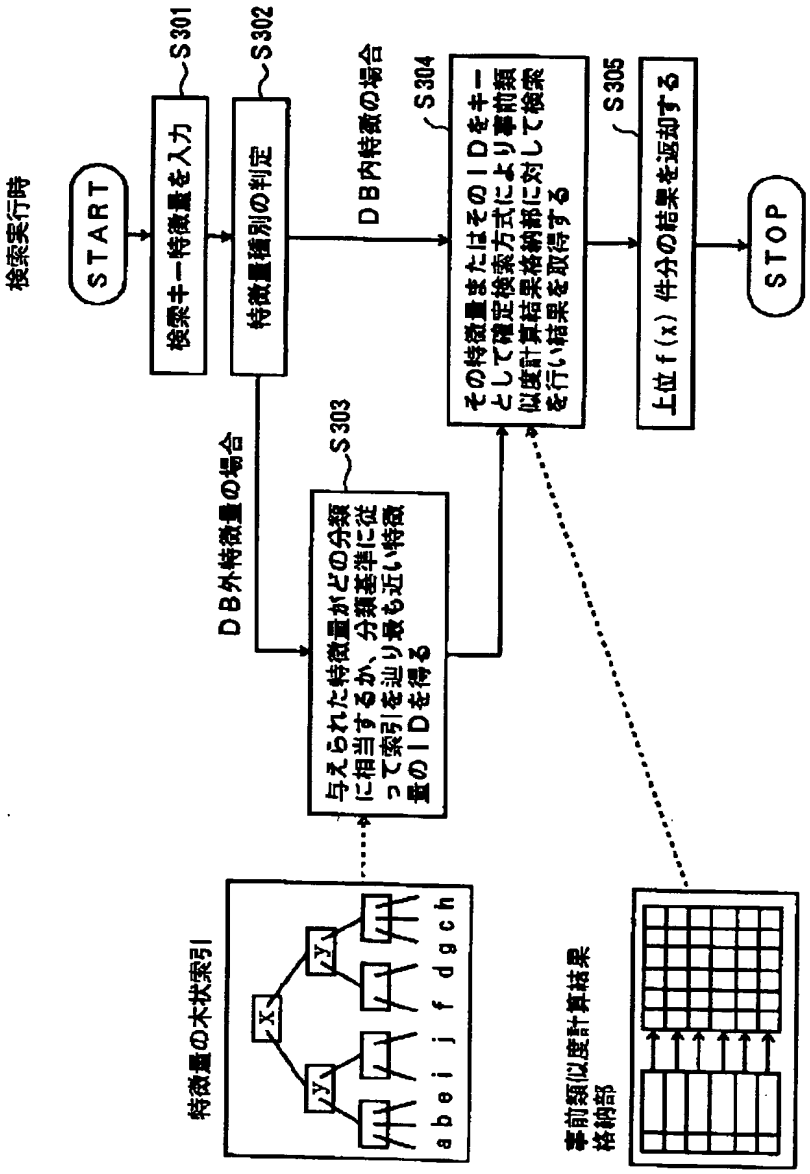
【図４】

本発明の索引構築時及び検索実行時の動作を示すフローチャート



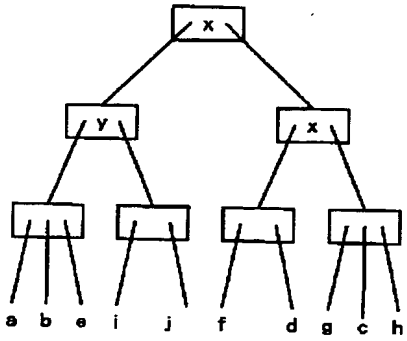
【図５】

本発明の最近傍検索の処理を示すフローチャート



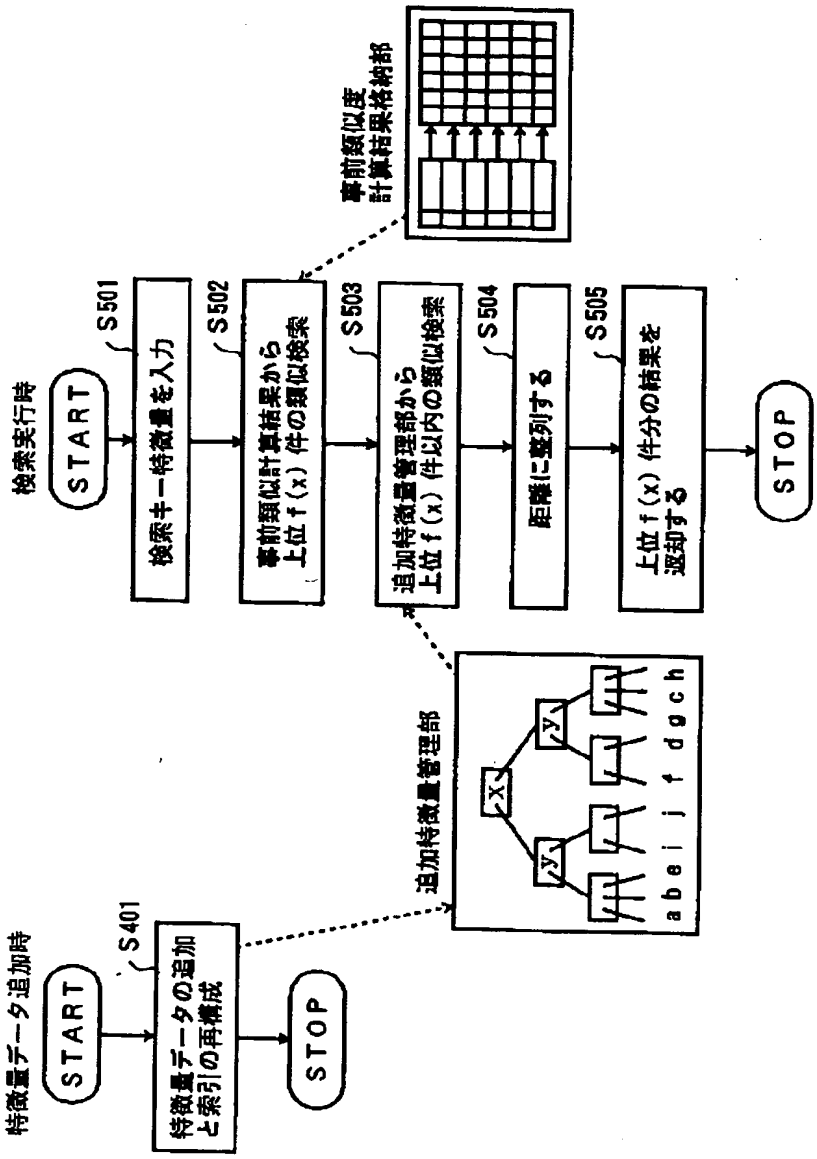
【図１２】

従来のＲ－ｔｒｅｅの木構造を説明するための図



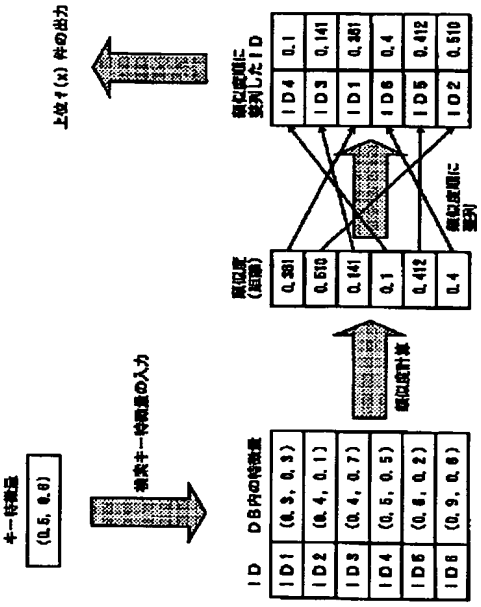
【図６】

本発明の特徴量データの追加がある場合の処理を示すフローチャート



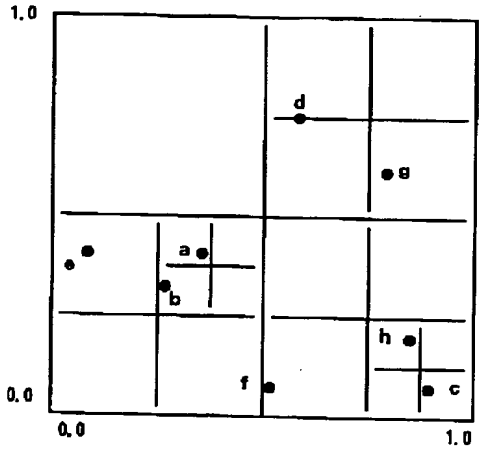
【図１０】

類似検索を説明するための図



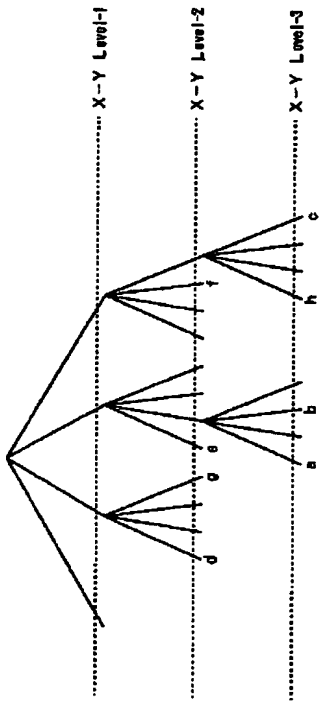
【図１３】

従来のPR－quadtreeの特徴量空間分割を説明するための図



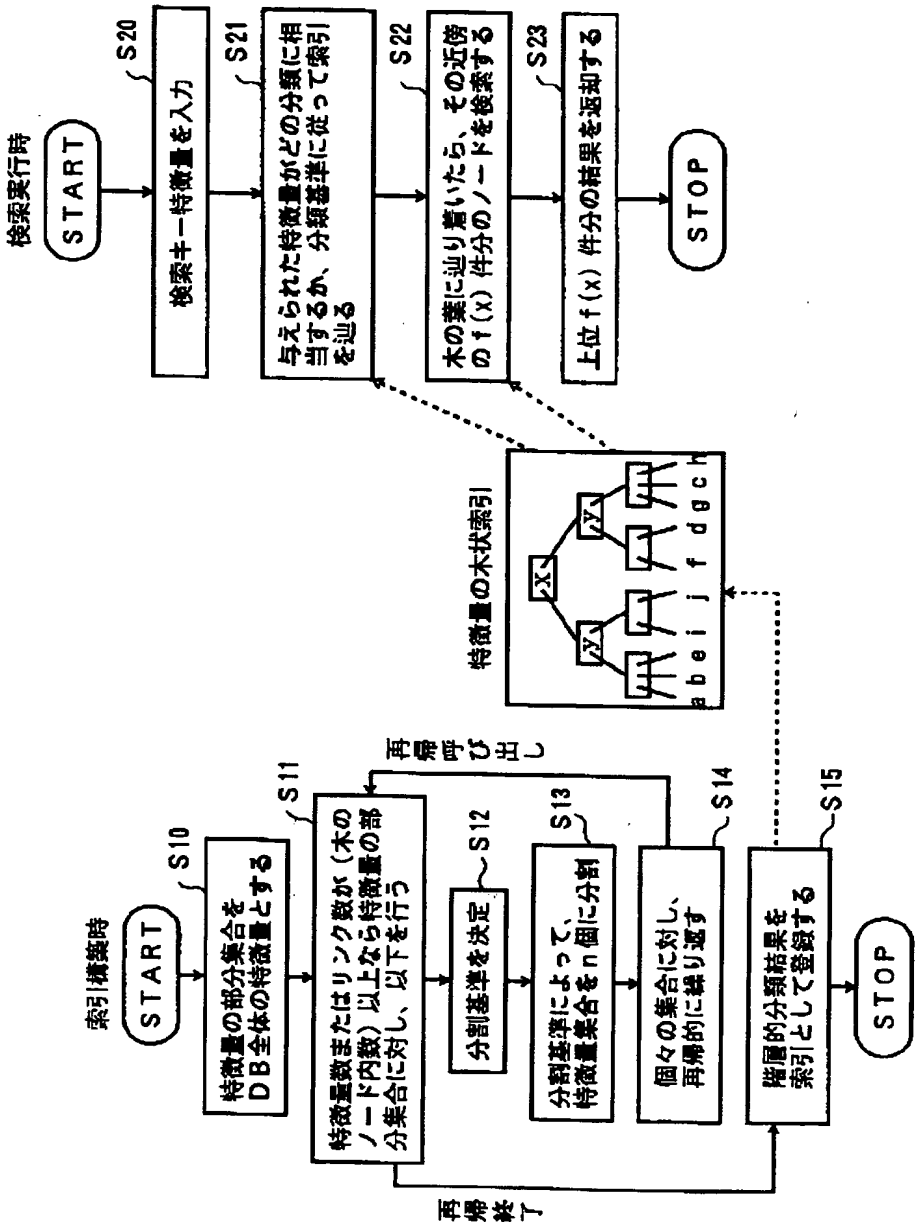
【図１４】

従来のＰＲ－ｑｕａｄｔｒｅｅの木構造を説明するための図



【図１５】

従来の木状索引を使った類似検索のフローチャート



フロントページの続き

(72)発明者　三井　一能
東京都新宿区西新宿三丁目19番２号　日本電信電話株式会社内

(72)発明者　串間　和彦
東京都新宿区西新宿三丁目19番２号　日本電信電話株式会社内

Ｆターム(参考)　5B050 EA24 FA10 GA08
5B075 ND07 ND12 ND14 ND40 PR06
UU13 UU40
5L096 JA04 KA09